Travaux Interdisciplinaires Parole et Langage (TIPA), 2000, n°19, 155-167.

MODE DE RECUEIL ET OUTIL D'ANALYSE D'UN CORPUS DE PAROLE SPONTANEE ETUDIE D'UN POINT DE VUE PSYCHOLINGUISTIQUE

Monique Vion & Annie Colas

Université de Provence

CNRS UMR 6057 : Parole et Langage, équipe Psycholinguistique

RESUME

L'article décrit d'abord les traits caractéristiques d'un corpus de parole spontanée, mis au service d'une approche psycholinguistique du développement des capacités narratives. Les narrations ont été recueillies auprès de 255 locuteurs francophones (des enfants de 7, 9 et 11 ans et des adultes), dans des conditions de production contrôlées expérimentalement. Il présente ensuite les objectifs et les points forts du système automatisé choisi pour explorer les productions. Il expose brièvement les contraintes de transcription nécessaires à l'édition dans le mode CHAT (Codes for the Human Analysis of Transcripts) du système CHILDES (Child Data Exchange System) ainsi que les principales fonctions de recherche offertes par le mode CLAN (Computerized Language Analysis).

Mots—clés : computerized exchange system for language data, narrations.

Les questions que nous abordons par l'analyse d'un corpus de parole spontanée concernent le caractère incrémental de la production verbale. Dans les années 80, un certain nombre d'écrits en psycholinguistique (Kempen, 1977, 1978 ; Kempen & Hoenkamp, 1982, 1987) ont souligné le caractère incrémental de la production verbale orale dont a tenu compte le modèle du système de la production à architecture modulaire défendu notamment par Levelt (1989) et Bock (1995). Dans ce modèle du système de traitement du

locuteur, la production d'un énoncé est conçue comme guidée par un message, à savoir, par une structure conceptuelle, qui se présente comme une séquence cumulée et ordonnée de fragments de contenu. Chaque fragment est traité d'étage en étage par les différentes composantes du système de traitement, depuis sa conception jusqu'à son articulation. Dès qu'un fragment conceptuel est disponible, il est transmis à la composante grammaticale. La composante grammaticale le traduit dans un fragment de phrase. Ce fragment est ensuite articulé. Pendant ce temps, le travail se poursuit sur d'autres fragments conceptuels et syntaxiques. De sorte que, toutes les composantes travaillent en parallèle, mais sur différents fragments d'information de l'énoncé en cours d'élaboration. Sur le schéma de la figure 1, le décalage en cascade des liens montre comment circule l'information au cours du temps.

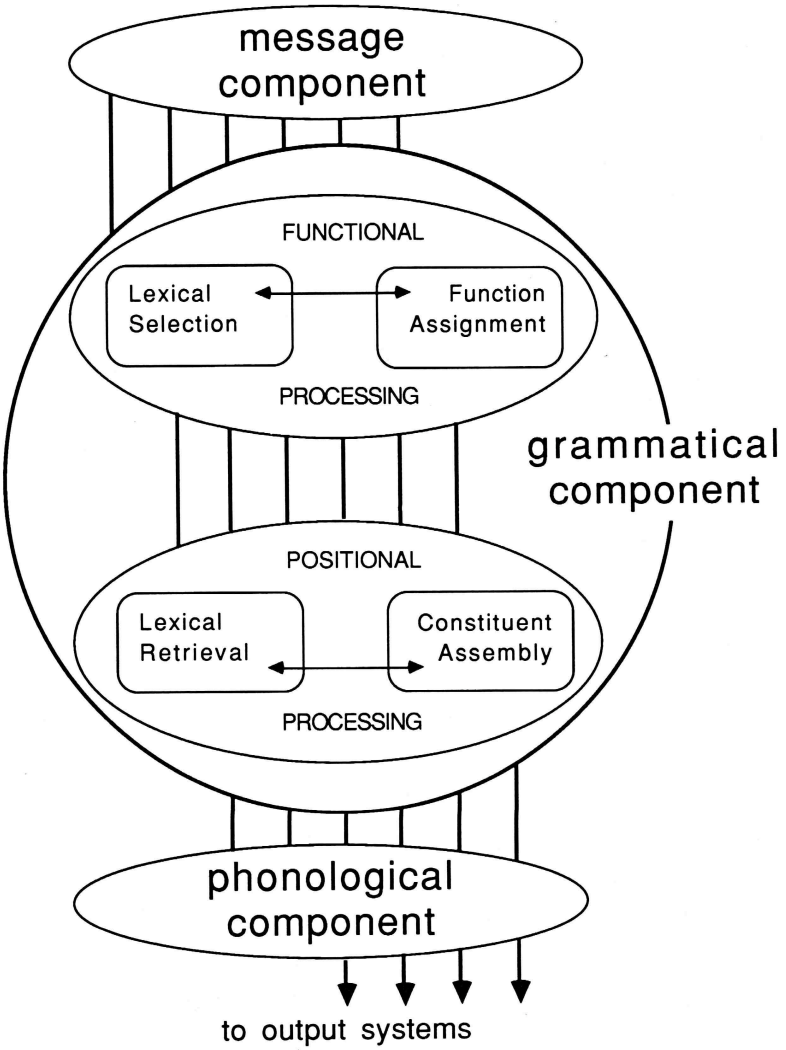


Figure 1 : Les composantes du système de production du langage (extrait de Bock, 1995).

Notre intérêt se focalise sur le passage de l'information de la composante conceptuelle à la composante grammaticale. Plus précisément, sur la façon dont le contenu d'un message est planifié et mis en mots pour être présenté au destinataire. A ce niveau du traitement, l'élaboration conceptuelle (la représentation pré-linguistique du contenu) est conçue comme ce qui sert de plan pour les processus de formulation linguistique.

L'hypothèse générale que nous tentons de vérifier est que les expressions linguistiques qui structurent le discours sont la manifestation de contraintes de nature conceptuelle inhérentes à la gestion de l'information (Bronckart, 1985, Chafe, 1986). Et que ces contraintes peuvent se montrer plus ou moins

propices à la mobilisation des compétences linguistiques au cours du développement de la vie entière.

Le but est donc d'intervenir sur les contraintes conceptuelles (cognitives) par la manipulation expérimentale, en contrôlant les conditions de production de façon à vérifier, par l'analyse du corpus recueilli, que ces contraintes déterminent les choix linguistiques des locuteurs.

Pour observer les différences liées au développement des compétences linguistiques et communicatives, les enregistrements de parole sont réalisés selon la méthode transversale. Celle-ci consiste à recueillir au même moment dans le temps pour les comparer, des observations auprès de groupes de personnes d'âge différent.

Structure du corpus

On a invité des locuteurs (enfants et adultes francophones) à raconter des bandes dessinées (BD) sans texte à un pair d'âge silencieux qui ne les connaissait pas. La consigne posait une obligation explicite : il s'agit d'une histoire et elle est racontable. De plus, la consigne demandait de rapporter le contenu des BD aussi fidèlement que possible, à savoir, en tenant compte de chaque image, mais en évitant cependant de donner trop de détails. Il était demandé au destinataire d'être un auditeur attentif, mais passif.

Le principe de construction des BD est illustré par la figure 2. La première image représentait toujours deux personnages. Les 7 images suivantes comportaient un seul des deux personnages engagé dans diverses actions. Dans un cas, les images suivant la première étaient, jusqu'à la dernière incluse, centrées sur un même personnage (continuité thématique). Dans l'autre cas, les images qui suivaient la première étaient centrées sur un même personnage jusqu'à l'avant-dernière. Mais la dernière image montrait ce que fait l'autre personnage de la première image (introduction d'une discontinuité thématique). Afin d'éviter de biaiser les productions recueillies par l'éventuelle saillance de

l'un des deux personnages due à sa position dans l'image, la disposition (gauche - droite) des personnages dans la première vignette a été contrebalancée.

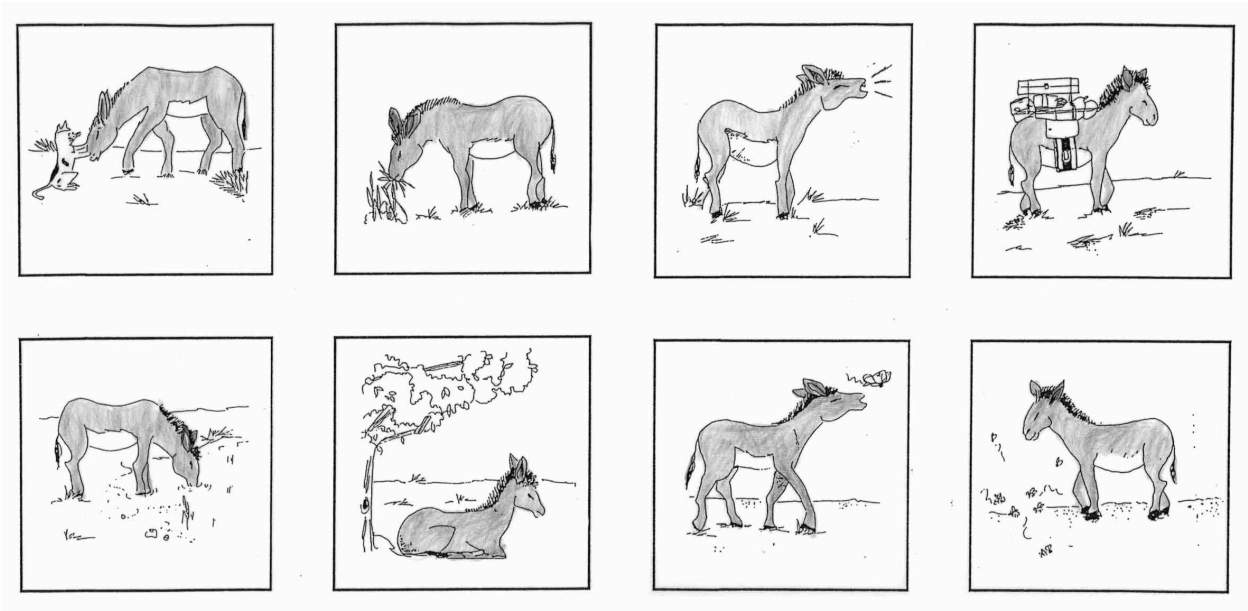


Figure 2a : Principe de construction des quatre versions d'une bande dessinée à deux personnages (thème maintenu).

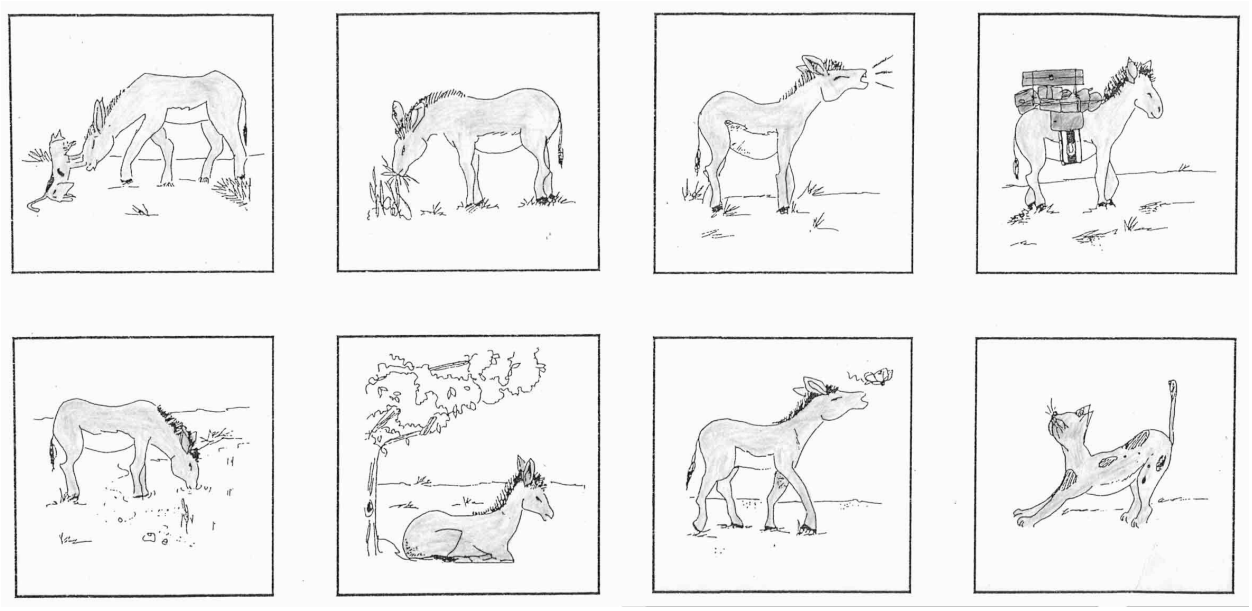


Figure 2b : Principe de construction des quatre versions d'une bande dessinée à deux personnages (thème changé).

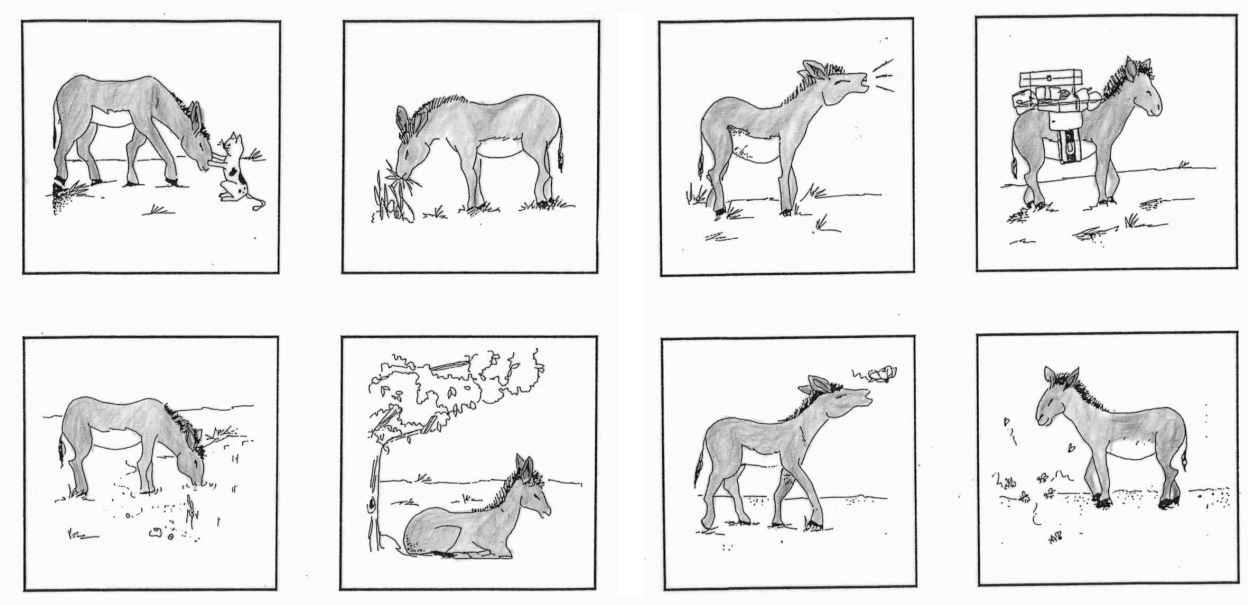


Figure 2c : Principe de construction des quatre versions d'une bande dessinée à deux personnages (thème maintenu).

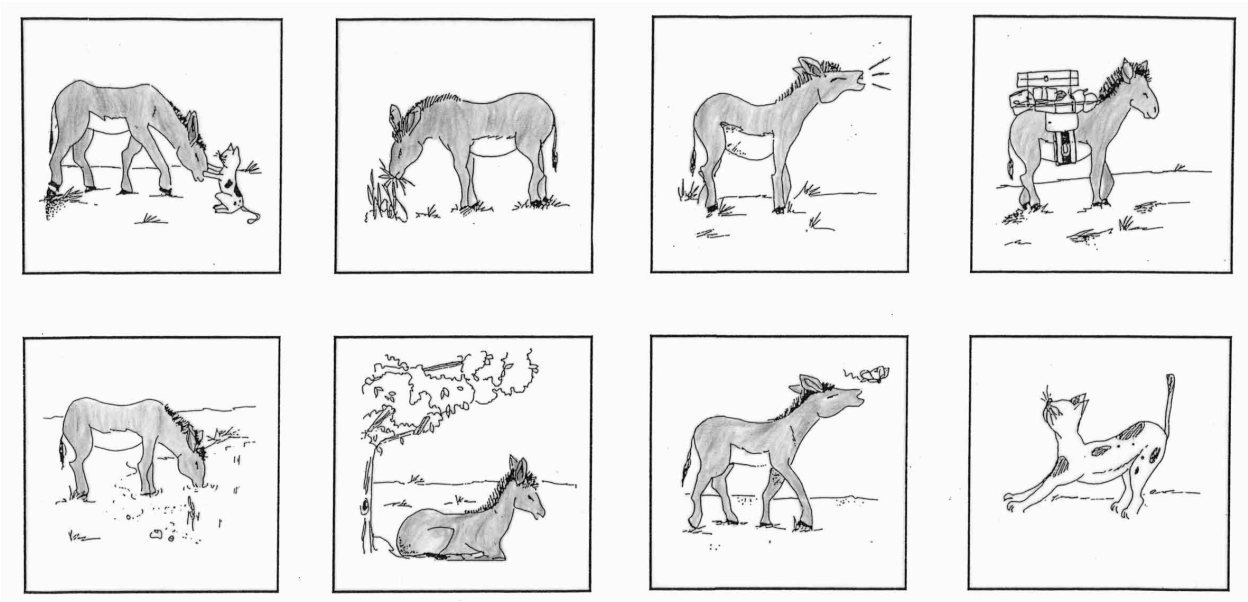
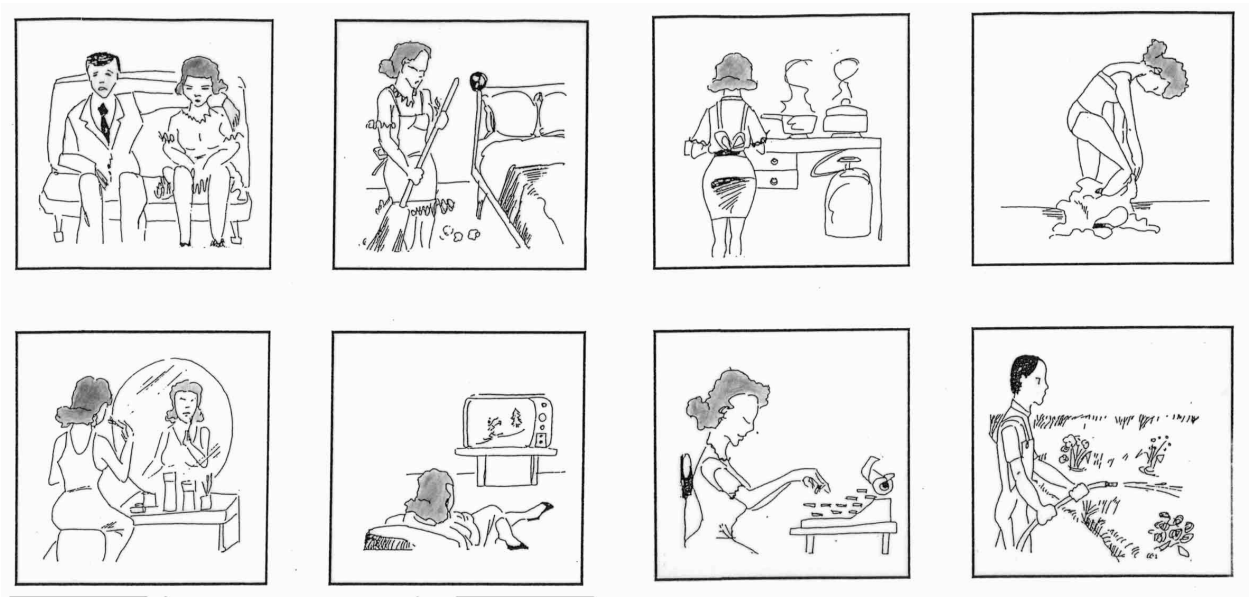


Figure 2d : Principe de construction des quatre versions d'une bande dessinée à deuxpersonnages (thème changé).

D'un point de vue cognitif, de nombreuses études ont montré que la narration sur la base d'images mobilise une activité d'inférence (voir les synthèses de Trabasso et Stein, 1997 ; Stein et Albro, 1997 sur ce point). Le locuteur doit non seulement comprendre les événements représentés sur chaque image, mais il doit aussi comprendre comment ces événements sont interconnectés. Il faut donc qu'il infère un contenu pour chaque image et qu'il se

construise une représentation d'ensemble de l'histoire en établissant des liens temporels et causaux entre les événements. La première contrainte cognitive manipulée a donc concerné l'activité d'inférence. Deux types de BD ont été contrastés qui s'opposent selon le caractère plus ou moins explicité par l'image de l'enchaînement des images successives (figures 3). Dans un cas, les événements, bien qu'organisés en séquence, sont interchangeables du point de vue de leur chronologie. Il est du ressort du locuteur d'inférer un enchaînement pour les images. Dans l'autre cas, les événements ne sont pas interchangeables du point de vue de la chronologie.



Figures 3a : Types d'enchaînement (arbitraire).

Figures 3b : Types d'enchaînement (ordonné).

La deuxième contrainte cognitive manipulée concernait la structuration du contenu. Dans un cas, la bande dessinée se présentait comme sur une page d'album. Les événements objets de la narration étaient accessibles en permanence. La prospection et la rétrospection étaient donc possibles. Dans l'autre cas, la bande dessinée était présentée sous la forme d'un carnet à raison d'une image par page. Le locuteur devait verbaliser "en ligne", au fur et à mesure qu'il prenait connaissance des images. De la sorte l'empan de planification du locuteur se trouvait limité.

Deux cent cinquante-cinq locuteurs (garçons et filles) ont participé au recueil. Soit 63 enfants de 7ans (âge médian : 6;6 ans), 64 enfants de 9 ans (âge médian : 8;8 ans), 64 enfants de 11 ans (âge médian : 10;6 ans) et 64 adultes étudiants. Chaque locuteur a été confronté à l'un ou l'autre mode de présentation des images (simultané *versus* consécutif), à l'un ou à l'autre type d'enchaînement (arbitraire *versus* ordonné) et aux quatre versions. Chaque locuteur a vu huit BD test à deux personnages (chacune étant présentée sous l'une des quatre versions) et trois BD de remplissage à un seul personnage (intercalées toutes les deux bandes dessinées test).

Dans un premier temps, les 2040 productions enregistrées (255 locuteurs x 8 BD) ont été intégralement transcrites selon les conventions de Hickmann, Hendriks, Roland et Liang (1994), mises au point pour étudier le développement de la cohésion du discours.

Sur ce corpus, nous avons étudié la façon dont les locuteurs avaient introduit les personnages dans le discours (Vion et Colas, 1998), géré la référence ultérieure aux mêmes personnages (Vion et Colas, 1999a et 1999b) et marqué explicitement les liens inter-clausaux (Vion et Colas, 2000 ; Vion et Colas, sous presse). Très vite, le recours à un moyen de traitement automatique des données s'est avéré nécessaire. Nous avons utilisé le système CHILDES

(Child Data Exchange System), mis au point dans le domaine de l'acquisition du langage pour analyser les interactions verbales et non verbales des dyades adulte-enfant (MacWhinney, 1991).

Le système CHILDES

Le système CHILDES a trois objectifs principaux : 1) apporter une précision scientifique au recueil des données, à la transcription de ces données et au codage, 2) automatiser l'analyse et 3) élargir une base de données empiriques. Pour cela, trois outils distincts et cependant intégrés ont été développés. Le mode CHAT qui regroupe les conventions pour transcrire les données et les coder. Le mode CLAN qui permet un grand nombre d'analyses automatiques sur les données transcrites et enrichies. Enfin le système compile et met à jour une base qui réunit les corpus recueillis depuis les années 60 en vue d'étudier l'acquisition du langage.

CHAT (Codes for the Human Analysis of Transcripts)

Le recours à CHILDES suppose de constituer des fichiers (un par locuteur ou par dyade) et de se soumettre aux règles d'écriture propres au mode CHAT. CHAT comporte trois niveaux d'écriture : les en-têtes de fichiers, les lignes principales et les lignes de codage.

Les en-têtes de fichier sont introduits par : « @ ». Ils définissent le fichier. Les en-têtes sont constants et obligatoires. Ils permettent de déclarer les informations de début (@begin) et de fin de fichier (@end). Ils servent aussi à introduire les informations sur le (ou les) participant(s). Les lignes principales sont introduites par : « * » suivie des 3 initiales du nom du locuteur. Elles comportent la transcription des énoncés du locuteur. Les lignes dépendantes sont introduites par : « % ». Ce sont les lignes de codage. CHAT offre des dimensions prédéfinies d'analyse (expressions faciales, intonation, syntaxe, phonologie, actes de langage, etc.) et des codes (prédéfinis ou bien laissés au

choix du chercheur). Ces lignes, introduites par %, ont des noms réservés de trois caractères qui sont déclarés dans le système (Tableau 1).

Tableau 1 : Liste des lignes dépendantes.

Symbol	*Description*
%act:	actions
%add:	addressee
%alt:	alternative transcription
%cod:	general purpose coding
%com:	comments by investigator
%def:	codes from SALT
%eng:	English translation
%err:	error coding
%exp:	explanation
%fac:	facial actions
%flo:	flowing version
%gls:	target language gloss for unclear utterance
%gpx:	gestural and proxemic activity
%int:	intonation
%mod:	model or target phonology
%mor:	morphemic semantics
%par:	paralinguistics
%pho:	phonetic transcription
%sit:	situation
%spa:	speech act coding
%syn:	syntactic structure notation
%tim:	time stamp coding

CLAN (Computerized Language Analysis)

Pour reconnaître des formes dans les lignes principales ou des codes dans les lignes dépendantes et effectuer des statistiques descriptives (rapport type/occurrence, longueur moyenne des énoncés, fréquence, analyse de co-occurrence, d'interactions, etc.), pour un locuteur ou pour plusieurs, sur tout ou partie d'un fichier CLAN propose un certain nombre d'outils (ou fonctions) dont certains sont décrits dans le tableau 2.

Tableau 2 : Quelques fonctions de CLAN.

CHAINS	Permet de suivre les séquences de codes entre les participants.
CHIP	Permet une analyse de l'interaction. Il compare deux énoncés particuliers et produit une analyse qui est introduite sous forme d'un nouveau « codetier » introduit par %. Le premier énoncé dans la paire d'énoncés est considéré comme l'énoncé-source et le second comme l'énoncé-réponse. La réponse est comparée à la source.
CHSTRING	Permet la modification de chaînes de caractères dans un fichier, on peut ainsi préserver l'anonymat en changeant les noms de manière automatique.
COLUMNS	Dispose CHAT en colonnes. Il permet de changer le format du protocole en mettant les productions d'un participant dans l'une et celles de l'autre participant dans l'autre par ex.
COMBO	Effectue une recherche booléenne sur des groupes de lignes et non sur des lignes. Très utile

	pour l'analyse syntaxique.
COOCCUR	Analyse des cooccurrences. Ce programme fait des tabulations pour la cooccurrence de mots (utile pour l'analyse de groupes syntaxiques tabulation).
DATES	Calcule l'âge à partir des dates.
DIST	Distance entre les codes. Ce programme permet de calculer combien d'énoncés existent entre l'occurrence spécifiée et un mot clé et un code.
FREQ	Comptage des occurrences.
GEM	Marquage de passages intéressants.
MTT	Longueur moyenne des tours de parole.
MLU	Longueur moyenne des énoncés.
MODREP	Appariement des mots d'une ligne dépendante à une autre.
MOR	Analyse morphologique.
PHONFREQ	Analyse de la fréquence phonémique.

Pour notre part, nous avons utilisé les fonctions MLU, FREQ et GEM. La fonction GEM permet de sélectionner des passages intéressants dans le corpus et de faire sélectivement des statistiques descriptives sur ces passages. L'encadré 1 présente en exemple une partie de fichier prêt pour nos analyses.

(1) exemple extrait d'un fichier codé.

```
@begin
@participants:	CHI	Coralie	child
@ID:	110.07.lm.s
.
.
.
@G:	un homme et une femme
@bg:	V1m
*CHI:	papa et maman sont tristes.
%flo:	aETs
*CHI:	i sont sur le canapé, entrain de réfléchir, comment i peut avoir un
  bébé.
@eg:	V1m
@bg:	VUm
*CHI:	maman balaye, balaye, et passe son temps à faire son lit, pour
  trouver comment elle va avoir un bébé.
%flo:	aETpy
*CHI:	ensuite, elle fait la cuisine et réfléchit, comment elle va avoir un
  bébé.
%flo:	aETv
*CHI:	elle se met son tutu, pour aller à la danse.
*CHI:	ensuite, après sa danse, elle mange au restaurant.
*CHI:	elle se maquille, bien, et elle se regarde dans la glace.
%flo:	aETprs
*CHI:	+^ après elle a fini sa danse, elle va regarder la télé.
*CHI:	elle voit jeux olympiques, qui font du ski, elle se demande, elle se
  dit +"/.
*CHI:	+" oh j'ai beaucoup de choses à faire, j'ai envie d'avoir un bébé, si
  je faisais du ski?
*CHI:	elle tape à la machine à écrire tout en réfléchissant.
*CHI:	elle dit +"/.
*CHI:	+" oh mince je me suis trompée, j'étais entrain de
  réfléchir si je pouvais avoir un bébé.
*CHI:	elle efface ça.
@eg:	VUm
@bg:	V8m
*CHI:	+^ elle va au marché, tout en réfléchissant encore.
*CHI:	et dit au monsieur +"/.
%flo:	aETd
*CHI:	+" hé monsieur je pourrai acheter ça?
*CHI:	+" mais est+ce+que vous avez une idée comment je pourrai avoir un
  bébé ?
@eg:	V8m
.
.
.
@end
```

Les usagers du système CHILDES sont informés des mises à jour régulières du système. Récemment, des connexions ont été établies entre ce système et le système d'analyse de la parole PRAAT (Boersma, & Weenink, 1992-2000) en usage au Laboratoire Parole et Langage. Ceci nous donne bon espoir d'identifier sur la base de notre corpus certaines des contraintes auxquelles obéit la planification des unités intonatives.

Références

Bock, K. (1995). Sentence production : from mind to mouth. In J.L. Miller et Eimas, P.D. (Eds.). *Speech Language and Communication* (pp.181-216). London: Academic Press.

Boersma, P., & Weenink, D. (1992-2000): Praat, a system for doing phonetics by computer. Available http://www.praat.org.

Bronckart, J.P. (1985). Pour un modèle de production du discours. In J.P. Bronckart (Ed.), *Le fonctionnement des discours* (pp 3-58). Neuchâtel : Delachaux et Nietslé.

Chafe, W. (1986). Cognitive constraints on information flow. In R. Tomlin (Ed.*) Coherence and grouding in discourse. Typological studies in language*. Vol. 11 (pp 21-51). Amsterdam : john Benjamins Publishing Compagny.

Hickmann, M., Hendriks, H., Roland, F., & Liang, J. (1994). *The development of reference to person, time and space in discourse : A coding manua*l. Nijmegen: Max Planck Institute for Psycholinguistics.

Kempen, G. (1977). Conceptualizing and formulation sentence production. In S. Rosenberg (Ed.), *Sentence production : Developments in research and theory*. Hillsdale, NJ : Erlbaum, 259-274.

Kempen, G. (1978). Sentence construction by a psycholgically plausible formulator. In R. Campbell & P. Smith (Eds.), *Recent advances in the psychology of language* : Vol. 2. Formal and experimental approaches. New York : Plenum Press, 103-123.

Kempen, G., & Hoenkamp, E. (1982). Incremental sentence generation : implication for the structure of a syntactic processor. In J. Horecky (Ed.), *Proceedings of the ninth International Conference on computational linguistics*. Amsterdam : North Holland.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive science, 11*, 201-258.

Levelt, J.M. (1989). *Speaking. From intention to articulation*. London: MIT Press.

MacWhinney, B. (1991). *The Childes Project. Tool for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum. Available FTP: Hostname : childes.psy.cmu.edu.

Stein, N., & Albro, E. (1997). Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In M. Bamberg (Ed.), *Narrative development: Six approaches* (pp.5-44). Mahwah, NJ: Lawrence Erlbaum.

Trabasso, T. & Stein, N. (1997). Narrating, representing and remembering event sequences. In P. van den Brœk, P., Bauer, & T. Bourg (Eds.). *Developmental spans in event comprehension and representation, Bridging fictional and actual events* (pp.237-270). Mahwah, NJ: Lawrence Erlbaum.

Vion, M., & Colas A. (1998). L'introduction des référents dans le discours en français : contraintes cognitives et développement des compétences narratives. *L'Année Psychologique*, *98*, 37-59.

Vion, M., & Colas, A. (1999a) Maintaining an reintroducing referents in French : cognitive constraints and development of narrative skills. *Journal of Experimental Child Psychology, 72*, 32-50.

Vion, M., & Colas, A. (1999b) Expressing coreference in French : cognitive constraints and development of narrative skills. *Journal of Psycholinguistic Research, 28*, n°3, 261-291.

Vion, M., & Colas, A. (juin 2000). Using connectives in French : cognitive constraints and development of narrative skills (le cas de la narration de séquences ordonnées d'événements). Communication : *6ème Congrès international de l'ISAPL* (International Society of Applied Psycholinguistics), Caen.

Vion, M., & Colas, A. (2000). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements*). Proceedings of the 8th Conference of the International Association for the Study of Child Language*, San Sebastian, Spain.

RESUME

L'article décrit d'abord les traits caractéristiques d'un corpus de parole spontanée, mis au service d'une approche psycholinguistique du développement des capacités narratives. Les narrations ont été recueillies auprès de 255 locuteurs francophones (des enfants de 7, 9 et 11 ans et des adultes), dans des conditions de production contrôlées expérimentalement. Il présente ensuite les objectifs et les points forts du système automatisé choisi pour explorer les productions. Il expose brièvement les contraintes de transcription nécessaires à l'édition dans le mode CHAT (Codes for the Human Analysis of Transcripts) du système CHILDES (Child Data Exchange System) ainsi que les principales fonctions de recherche offertes par le mode CLAN (Computerized Language Analysis).

SUMMARY

This paper describes, first, the characteristics of a spontaneous speech corpus which deals with a psycholinguistic approach of the development of narrative skills. Narratives were produced by 255 French—speaking participants (7—, 9— and 11—yr—old children and adults), in experimentally controlled conditions of production. Second, the paper presents the purposes and the main characteristics of the automatized system choosen to explore the productions. We briefly show the constraints of transcription needed for the edition of protocols in CHAT mode (Codes for the Human Analysis of Transcripts), and the main search functions of CLAN mode (Computerized Language Analysis).

Key words : computerized exchange system for language data, narratives.